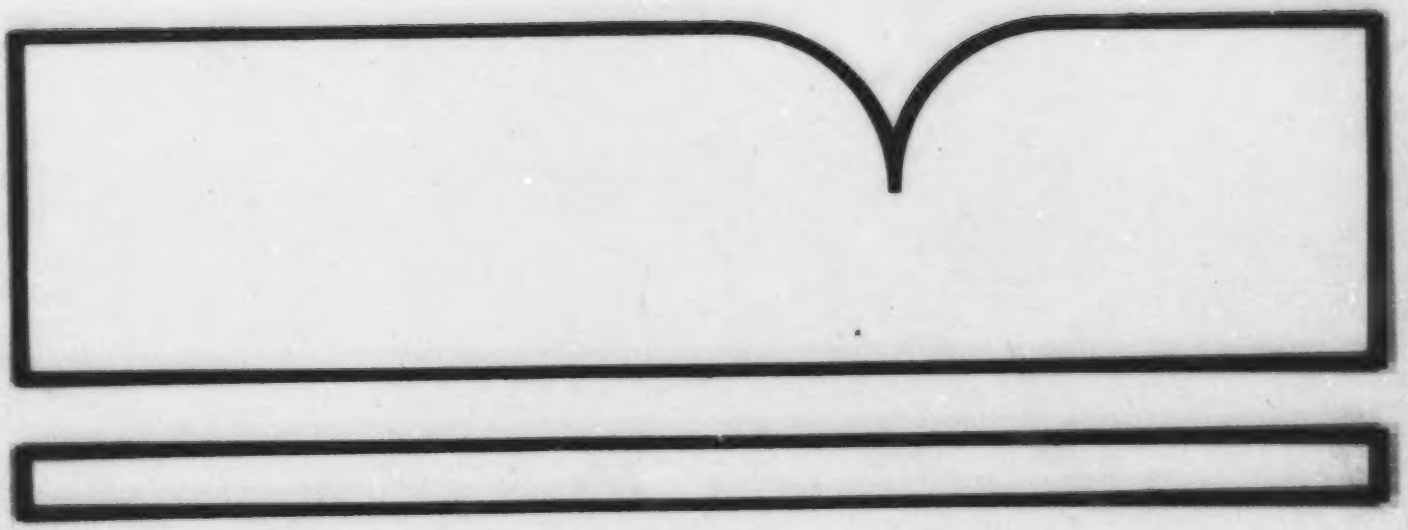
PB85-189736

# Trace Organic Compound Identification Shannon Information Content and Pattern Recognition of Mass Spectra

(U.S.) Environmental Monitoring Systems Lab.
Research Triangle Park, NC

1985



U.S. Department of Commerce
National Technical Information Service
NTIS

PB85-189736

EPA/600/D-85/066
1985

# TRACE ORGANIC COMPOUND IDENTIFICATION: SHANNON INFORMATION CONTENT AND PATTERN RECOGNITION OF MASS SPECTRA

by

Donald R. Scott
Environmental Monitoring Systems Laboratory
U.S. Environmental Protection Agency
Research Triangle Park, NC 27711

ENVIRONMENTAL MONITORING SYSTEMS LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
RESEARCH TRIANGLE PARK, NC 27711

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

**TECHNICAL REPORT DATA**
*(Please read Instructions on the reverse before completing)*

| 1. REPORT NO. | 2. | 3. RECIPIENT'S ACCESSION NO. |
|---|---|---|
| EPA/600/D-85/066 | | |

| 4. TITLE AND SUBTITLE | 5. REPORT DATE |
|---|---|
| Trace Organic Compound Identification: Shannon Information Content and Pattern Recognition of Mass Spectra | 1985 |
| | 6. PERFORMING ORGANIZATION CODE |
| 7. AUTHOR(S) | 8. PERFORMING ORGANIZATION REPORT NO. |
| Donald R. Scott | |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS | 10. PROGRAM ELEMENT NO. |
| U. S. Environmental Protection Agency<br>Environmental Monitoring Systems Laboratory<br>Methods Development and Analysis Division<br>Research Triangle Park, North Carolina 27711 | 11. CONTRACT/GRANT NO. |
| 12. SPONSORING AGENCY NAME AND ADDRESS | 13. TYPE OF REPORT AND PERIOD COVERED |
| Environmental Monitoring Systems Laboratory<br>Office of Research and Development<br>U.S. Environmental Protection Agency<br>Research Triangle Park, NC 27711 | 14. SPONSORING AGENCY CODE<br>EPA/600/08 |

15. SUPPLEMENTARY NOTES

16. ABSTRACT

The analytical process may be identified with a system containing three major components: sampling, analysis, and data interpretation. Each component may be used as a filter to eliminate a large number of organic compounds from further consideration. Optimization of the system requires maximum discrimination for each filter (component). The mass distribution and Shannon information content of low resolution, binary encoded mass spectra have been calculated for a set of 78 volatile organic compounds which are routinely sought in ambient air samples. This set of compounds and their binary encoded mass spectra will be used as an example of preselection of variables for SIMCA pattern recognition using Shannon information content. A total of 153 different masses in the spectra of the 78 compounds are compressed into a set of 16 key masses. This set of 16 masses contains a maximum of 13 bits of information, neglecting correlation between masses. These 13 bits of information theoretically are sufficient to distinguish among 8200 compounds. The Shannon information content for different analytical methods also will be discussed.

17. KEY WORDS AND DOCUMENT ANALYSIS

| a. DESCRIPTORS | b. IDENTIFIERS/OPEN ENDED TERMS | c. COSATI Field/Group |
|---|---|---|
| | | |

| 18. DISTRIBUTION STATEMENT | 19. SECURITY CLASS *(This Report)* | 21. NO. OF PAGES |
|---|---|---|
| RELEASE TO PUBLIC | UNCLASSIFIED | 11 |
| | 20. SECURITY CLASS *(This page)* UNCLASSIFIED | 22. PRICE |

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

# TRACE ORGANIC COMPOUND IDENTIFICATION: SHANNON INFORMATION CONTENT AND PATTERN RECOGNITION OF MASS SPECTRA

Donald R. Scott

U. S. Environmental Protection Agency
Environmental Monitoring Systems Laboratory
Research Triangle Park, North Carolina 27711

The analytical process may be identified with a system containing three major components: sampling, analysis, and data interpretation. Each component may be used as a filter to eliminate a large number of organic compounds from further consideration. Optimization of the system requires maximum discrimination for each filter (component). The mass distribution and Shannon information content of low resolution, binary encoded mass spectra have been calculated for a set of 78 volatile organic compounds which are routinely sought in ambient air samples. This set of compounds and their binary encoded mass spectra will be used as an example of preselection of variables for SIMCA pattern recognition using Shannon information content. A total of 153 different masses in the spectra of the 78 compounds are compressed into a set of 16 key masses. This set of 16 masses contains a maximum of 13 bits of information, neglecting correlation between masses. These 13 bits of information theoretically are sufficient to distinguish among 8200 compounds. The Shannon information content for different analytical methods also will be discussed.

## INTRODUCTION

Chemical analysis can be considered to be a system with three major components: a) sampling, b) analysis (i.e., separation and detection) and c) data interpretation. Each component of the system is important to the overall result, and quality assurance measures must be taken to ensure that all components are functioning properly. Sampling is extremely important since errors in sampling will affect the two subsequent components, analysis and data interpretation. Data interpretation, including data reduction and transformation, is the final step and perhaps the most important of the three. It is probably true that greater improvements in the chemical analysis system could be achieved at relatively low cost by improving data interpretation than by great expenditures of funds on the other two steps in the process.

Each of the system components can be considered to act as a filter to eliminate undesired compounds from further consideration. If properly designed, the sampling process itself can eliminate literally thousands of compounds in complex samples, such as ambient air. For optimization of the

total system, i.e., the chemical analysis process, each filter (component) must have maximum discrimination for the set of compounds sought. If incorrectly applied, the sampling and analysis steps can also add artifacts to the system. Since full control of all three components of the system is not always possible for optimization, concentration on the data interpretation component, which often may be manipulated at a later time, is perhaps justified. Pattern recognition, and other applied mathematical techniques, have great potential for optimizing the data interpretation component of the chemical analysis system. Transformation of the output from modern analytical instrumentation into the specific information required by the analyst for the specific problem at hand is the challenge for chemometric techniques.

In the present study, the information required to identify 78 volatile organic compounds in ambient air using low resolution mass spectra will be discussed. Binary encoded mass spectra of these compounds will be analyzed for their information content, and key masses for the identification of the set of compounds will be identified. The information content of various analytical techniques also will be briefly discussed.

## BINARY ENCODED SPECTRA AND SHANNON INFORMATION

It has been known for some time (1,2) that binary encoded spectra retain large portions of the information content of the respective original spectra. In binary encoded spectra the intensity measure in a given analysis channel is replaced by a 1, or 0, depending upon whether the intensity is above, or below, a specified threshold level. An analysis channel in low resolution mass spectrometry, as used in the present context, is a unit mass range. In infrared spectroscopy an analysis channel would be a specified wavelength, or frequency, range.

The Shannon information, which is related to the uncertainty (entropy) before and after analysis, can be expressed for binary encoded spectra (3-5) as

$$I \text{ per channel} = -p \log_2 p - (1-p) \log_2 (1-p).$$

Here p is the probability of occurrence of a given mass in a set of reference spectra of the compounds sought. This formula does not allow for encoding errors or for errors in the reference spectra, i.e., no uncertainty is assumed after analysis. The information content for each mass channel occurring in the set of compounds can be calculated from this formula. If correlations between masses are ignored, then a maximum information content for a specified set of masses may be obtained by summing the I for each mass channel over all the masses included. The unit of information, as used here, is the bit. The maximum I occurs at a probability of 0.5 and is one bit.

For binary encoded spectra, the information required for identification of one of N compounds is shown in Table I. For 100 compounds, 6.6 bits are required; while for 10000 compounds, 13.3 bits are required. The information shown in this table is a measure of the number of unique spectra in the reference spectra data base required to represent the desired number of compounds, which is given by $2^I$.

TABLE I

INFORMATION REQUIRED FOR COMPOUND IDENTIFICATION

For 1 of N Compounds:

| N | I (bits) | N | I (bits) |
|---|---|---|---|
| 10 | 3.3 | 10,000 | 13.3 |
| 100 | 6.6 | 50,000 | 15.6 |
| 500 | 9.0 | 100,000 | 16.6 |
| 1000 | 10.0 | 500,000 | 18.9 |
| 5000 | 12.3 | 1,000,000 | 19.9 |

TABLE II

ORGANIC COMPOUNDS SOUGHT

Number of Compounds: 78

Molecular Weight: 72 to 250

Median Molecular Weight: 126

Substituted Benzenes - 23

Halopropanes - 16

Haloethanes - 11

Halobutanes - 9

Halomethanes - 8

Haloethenes - 4

Halopropenes - 3

Halobutenes - 2

Cyclic Ethers - 2

## INFORMATION CONTENT OF MASS SPECTRA OF 78 COMPOUNDS

A set of 78 volatile organic compounds sought in ambient air samples was examined for information requirements for analysis and for the information content of their reference spectra in binary encoded form. These particular compounds are routinely sought in air samples in our laboratory and quantitatively determined by gas chromatography/mass spectrometry. A general description of the set of compounds is given in Table II. The molecular weights of the set range from 72 to 250 with a median of 126. Almost one-third of the compounds are substituted benzenes. The remainder are haloalkanes or haloalkenes with the exception of two cyclic ethers. The range of masses found in their mass spectra is from 35 to 256. Only 153 of these masses actually are found in the set of spectra. The empty channels occur at masses 66-68; 116; 143; 151-155; 177-178; 180-185; 187; 189; 191-199; 205; 207; 209; 211; 213; 215; 217; 219-249; 251; 253; and 255. These empty channels carry no information with regard to this set of compounds.

The information content of each occupied mass channel for this set of reference spectra was calculated from the distribution of the masses among the channels using a threshold level of 1% of the base peak in a spectrum. The base peak is the highest intensity peak in the spectrum of an individual compound. The results, which are shown in Figure 1, ranged from 0.099 to 1.00 bit per mass channel. The majority of the masses with high information content were less than mass 95. The 16 masses with the highest information content were selected and are listed in Table III. These key masses ranged from 41 to 107, and all had information contents of 0.80 bit or greater. The median information content was 0.91 bit. The key masses together contain a sum of 14.4 bits of information, neglecting errors and correlation effects. All 78 compounds have at least one of these masses in their individual mass spectra. Thirty-two of the 78 base peaks are included in this set of 16 masses. An estimated correction for errors (4) reduces the total information content to 13 bits. However, correction for correlation between the masses will certainly reduce the actual information to less than 13 bits. Therefore the 13 bits represent the maximum information content of this set of binary encoded mass spectra. One might suspect that these masses are only important for this particular problem. However, this set of key masses also contains a sum of 13 bits of information, neglecting correlation, with regard to a set of 9600 mass spectra (4). The 13 bits of information are sufficient to differentiate among 8200 different compounds.

## PATTERN RECOGNITION WITH KEY MASSES

One of the reasons for the identification of the 16 key masses was possibly to reduce the number of variables required in the SIMCA pattern recognition study of the mass spectra of these compounds being conducted by Professor Dunn of the University of Illinois at Chicago. These 16 masses could be used as a trial set for identification of the compounds. The reduction of the 153 different masses in the reference spectra to 16 key masses represents a considerable compression of the original set. To see how well

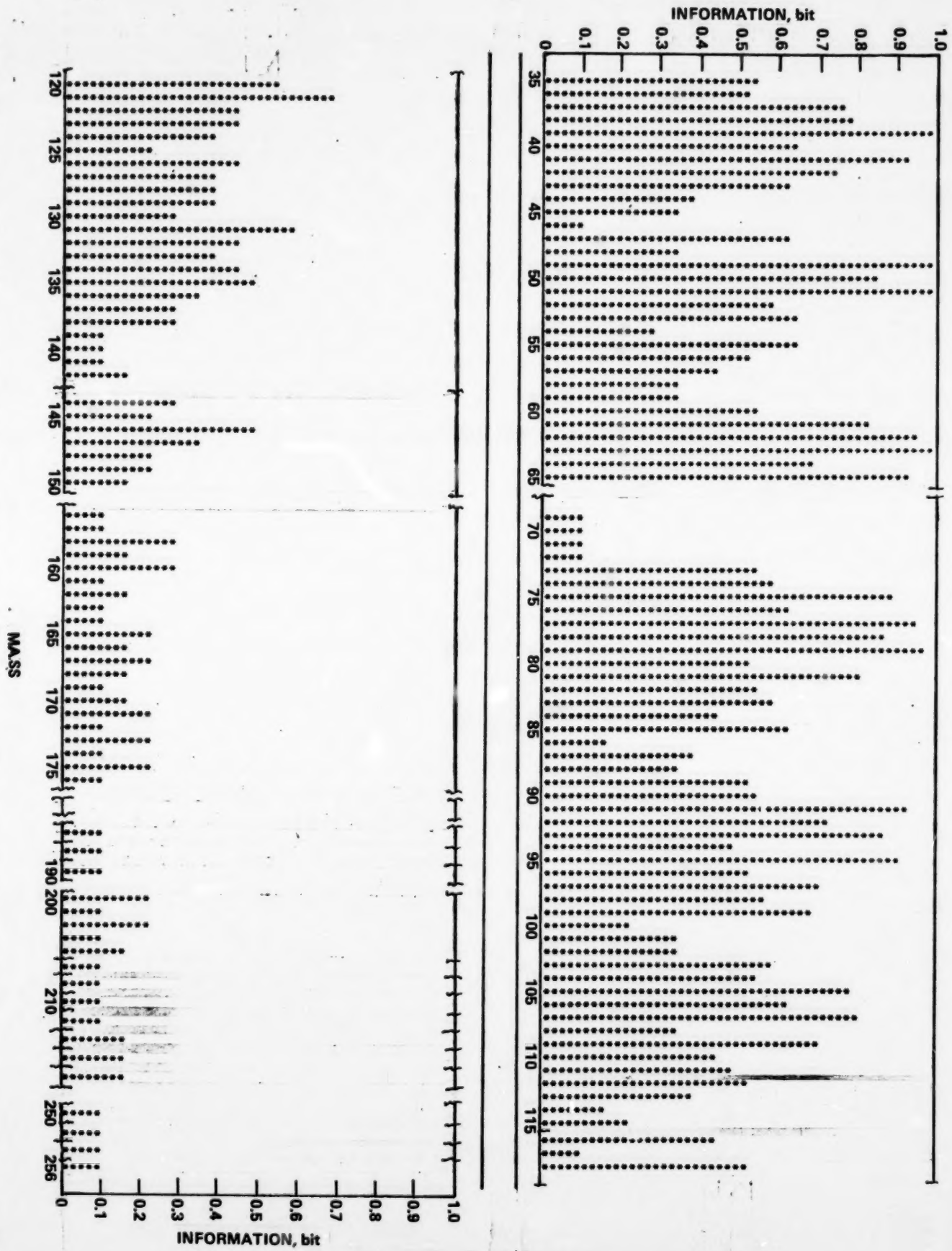



Figure 1. Maximum Information Content per Mass Channel for 78 Compounds

TABLE III

MOST INFORMATIVE MASSES FOR 78 COMPOUNDS[a]

| Mass | I (bits) | Σ I (bits) | No. of Additional Different Compounds[b] | Base Peaks[c] |
|---|---|---|---|---|
| 41 | 0.92 | | 26 | 5 |
| 51 | 0.99 | 1.91 | 22 | |
| 63 | 1.00 | 2.91 | 11 | 6 |
| 65 | 0.92 | 3.83 | | |
| 77 | 0.94 | 4.77 | | 2 |
| 79 | 0.96 | 5.73 | 6 | |
| 91 | 0.92 | 6.65 | 1 | 10 |
| 95 | 0.90 | 7.55 | 6 | |
| 49 | 0.98 | 8.53 | 1 | 1 |
| 50 | 0.84 | 9.37 | 4 | |
| 61 | 0.82 | 10.19 | 1 | 1 |
| 75 | 0.89 | 11.08 | | 4 |
| 78 | 0.84 | 11.92 | | 1 |
| 81 | 0.80 | 12.72 | | |
| 93 | 0.87 | 13.59 | | 1 |
| 107 | 0.82 | 14.41 | | 1 |
| Corrected for errors[d] | | 13 | Total: 78 | 32 |

[a] Selected on basis of high information content before correlation correction.
[b] Additional compounds having a given mass but not included with a previous mass.
[c] Number of compounds having a given base peak (highest intensity peak) of a given mass.
[d] Errors are estimated to be 0.09 bit per mass based on the lowest information content for any mass being set equal to zero.

the 16 masses differentiate among the 78 compounds, a SIMCA principal component analysis of the binary encoded mass spectra including only these 16 masses was performed on 49 of the compounds. Only 49 of the compounds could be examined due to limitations of the Osborne microcomputer and software used to perform the analysis. The results are shown in Figure 2, plotted as scores on principal component 1 vs. the scores on principal component 2. The total variance explained with the two principal components was 39% of the total. The plot shows that the use of the 16 mass binary encoded spectra provided for very good separation of these 49 compounds, with a few exceptions. Even though correlation between masses was not included in the information contents previously calculated, any correlations between the 16 masses would inherently be included in this plot. It appears, therefore, that the use of Shannon information content is an effective manner of selection of important variables for pattern recognition studies. This will be better understood when the complete SIMCA pattern recognition study has been completed using all 153 masses and all intensity data in the reference spectra.

## INFORMATION CONTENT OF SINGLE ANALYSIS METHODS

A summary of the information content of various single analytical methods taken from the work of Dijkstra et al. (3-5) is given in Table IV. Packed column gas chromatography has a relatively low information content of ca. 6 bits, which is sufficient to distinguish among only 64 different compounds. Mass spectrometry and infrared spectroscopy have 5 to 28 bits, depending on the number of channels used. For 20 channels both techniques have ca. 10 bits, which is sufficient to distinguish among 1020 compounds. This is a much larger number than is possible with gas chromatography, but it is not sufficient for some complex samples which require previous separation and/or the use of additional analytical techniques. Considering these results, one should be very careful in using gas chromatography alone for identification of organic compounds in complex mixtures. It also appears that infrared spectroscopy could be used for qualitative analysis more than it is presently. However, the detection limit and other limitations of a method are as important as the information content in selecting a method for trace organic identification.

## REFERENCES


1. S. L. Grotch, Anal. Chem. 42, 1214 (1970).
2. L. E. Wangen, W. S. Woodward, and T. L. Isenhour, Anal Chem. 43, 1605 (1971).
3. F. Dupuis and A. Dijkstra, Anal. Chem. 47, 379 (1975).
4. G. Van Marlen and A. Dijkstra, Anal. Chem. 48, 595 (1976).
5. P. F. Dupuis and A. Dijkstra, Z. Anal. Chem. 290, 357 (1978).



This paper has been reviewed in accordance with the U. S. Environmental Protection Agency's peer and administrative review policies and approved for presentation and publication.

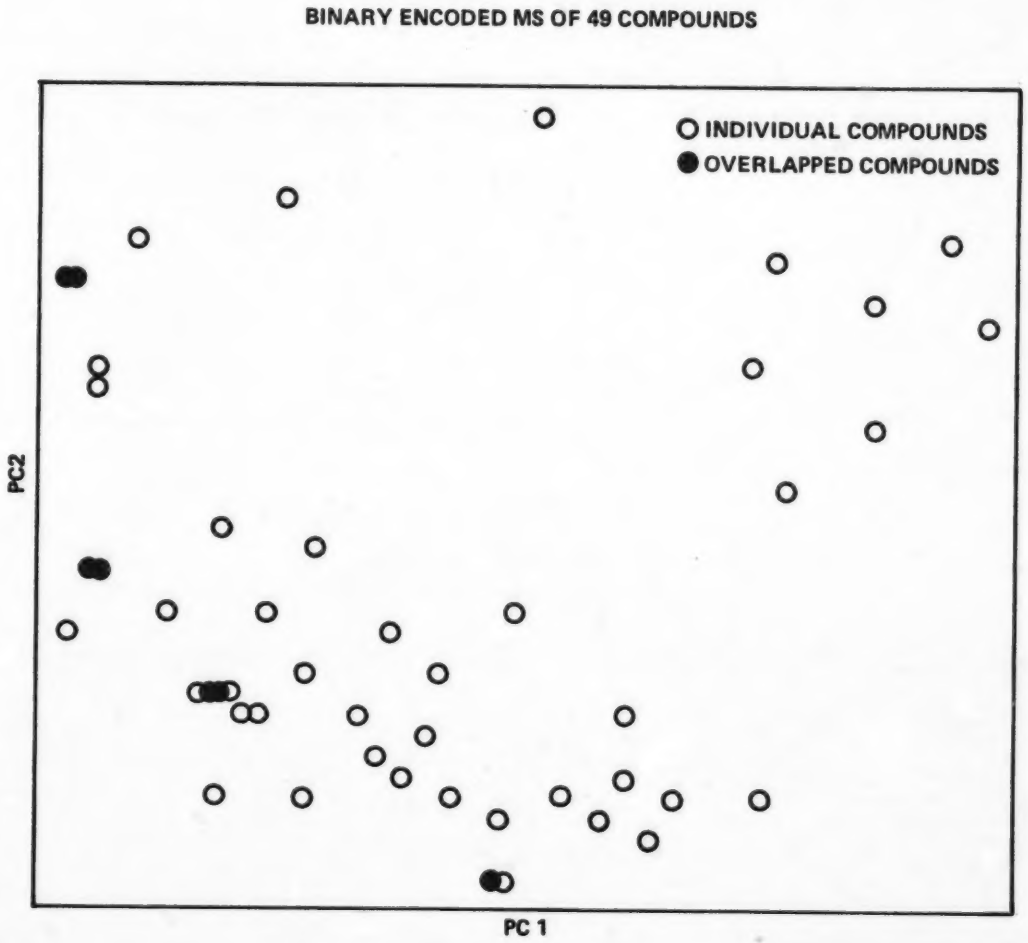



Figure 2. Principal Component Plot of Binary Encoded Key Mass Spectra of 49 Compounds

TABLE IV

INFORMATION CONTENT OF VARIOUS SINGLE METHODS

| | I (bits) | No. of Distinguishable Compounds |
|---|---|---|
| Gas Chromatography[a] | 6 | 64 |
| Low Resolution Mass Spectroscopy[b] | | |
| 10 peaks | 7 | 130 |
| 20 peaks | 12 | 4100 |
| 90 peaks | 28 | 270,000,000 |
| Infrared Spectroscopy[c] | | |
| 10 intervals | 5 | 32 |
| 20 intervals | 10 | 1020 |
| 90 intervals | 21 | 2,000,000 |

[a] Single packed column. 248 compounds. Reference 3.

[b] 10,000 binary encoded spectra corrected for correlations and errors. Reference 4.

[c] 97,000 binary encoded spectra corrected for correlations and errors. Reference 5.